Institute of Exercise Training and Sport Informatics,
German Sport University Cologne
Head of Institute: Prof. Dr Daniel Memmert

Data-driven performance analysis in soccer: A compilation of data science and machine learning techniques for pre-processing and knowledge discovery

Doctoral thesis approved for the degree
Ph.D. Exercise Science

by

Ashwin Ajay Phatak
from
Pune, India

Cologne, 2022

First reviewer: Prof. Dr. Daniel Memmert, Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Germany

Second reviewer: Prof. Dr. Jaime Sampaio, Universidade de Trás-os-Montes e Alto Douro Vila Real, Portugal

Chair of the doctoral committee: Prof. Dr. Mario Thevis, Institute of Biochemistry, German Sport University Cologne, Germany.

Thesis Defended on: 16.05.2023

Affidavits following §7 section 2 No. 4 and 5 of the doctoral regulations from the German Sport University Cologne, February 20, 2013:

Hereby I declare:
The work presented in this thesis is the original work of the author except where acknowledged in the text. This material has not been submitted either in whole or in part for a degree at this or any other institution. Those parts or single sentences, which have been taken verbatim from other sources, are identified as citations.

I further declare that I complied with the actual "guidelines of qualified scientific work" of the German Sport University Cologne.

Ashwin Ajay Phatak

01.08.2022

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Dr Daniel Memmert for giving me the opportunity and the freedom of to pursue a PhD at the German Sports University (GSU). The environment at the GSU was crucial for balancing the hard days of work with sports, so I would like to thank the GSU as a whole.

I would especially thank my peers and collaborators, without whom this work would have not been possible. I would like to thank Manuel for tolerating my jokes and opposing my wild ideas. Justus, for all the iron we pumped, podcasts we shared and for being the biggest geek I know. Marc, Max and Henrik for not only being an excellent work team but also to hold me up to a high standard. Fabi and Dominik for being excellent sounding boards. Robert for roasting us regularly, so we keep our feet on the ground (on behalf of the whole PhD room). I would also like to mention Marius, Phillip and Sebastian for bringing practical ideas to the otherwise nerdy data cave. A lot of the credit for my work is shared by Saumya and Mikael, who were great collaborators, actively helping me improve my research.

No endeavour in my life was or ever will be possible without the support of my family. Firstly my chosen brothers, Anton, Yago and Soma, everything from evening Chai tea, and salsa parties to emotional and physical support, anything I needed you were there. You were there the whole time, not just this work, but life in cologne would have not been possible without you. Also big thanks to all my close friends and the social dancing community in general for keeping me smiling and mentally healthy.

Last but not least, I would like to thank my father, mother, sister and cousin who were incredible mentors not only in life decisions but also providing valuable insights on a wide variety of technical topics relating to my work and business. It's rare to have a combination of talent and unconditional support from within the family, and I am thoroughly grateful for it.

Abstract

Big data has proven to be of increasing influence in a wide array of decision-making and knowledge-discovery processes across multiple domains. Gathering, pre-processing and analysis of data to make decisions or in the knowledge discovery process is a non-trivial task. With increasing access to computing power, a multitude of data science techniques have emerged as problem-solving tools in many domains of society. In the last decade, a vast amount of data has been collected in the field of sports. Several labs across the world have used this data in conjugation with Machine Learning/Data Science (ML/DS) techniques to add significant value to the sports industry and academics. However, compared to the potential of the available data, only a small subset has been exploited. This is primarily due to the lack of coding/programming expertise required to get the data in a form which is optimal for building models to answer specific questions of interest. The problem of pre-processing bottleneck has partially been solved due to data resources such as OPTA, STATSBomb and FBRef.com, which provide clean tracking, event and notational data. Additionally, libraries in Python and R such as Floodlight, AMIE, and SoccerAction offer packages which streamline pre-processing and visualization steps, thus offering great access to big data analysis for domain expertise with limited coding expertise. With these developments in mind, the current thesis aims to introduce ML/DS methods such as regression, binary classification, feature engineering and k-fold cross-validation into the field of sports analytics. This can potentially provide domain-specific experts, with the necessary technical tools to exploit the rising amount of data in the sports industry.

Through published case studies, each of which addresses a specific hypothesis. The thesis explains the importance of the normalization of KPIs as a feature engineering step before statistical modelling. It also elaborates on the value of using k-fold cross-validation as a model evaluation criterion for both regression and classification problems. The thesis further emphasises the value of using multiple ML models for solving specific problems as model robustness to avoid false findings due to the bias of a single algorithm. The provided methods can potentially be applied across research in general but the field of bat and ball sports like Cricket and Baseball seems to be conducive for big data analysis using ML. This is due to their unique closed-action nature (one action, one reaction leading to a result of that action-reaction pair) as they have a lower degree of randomness as compared to invasion sports. The thesis has a few limitations due to its scope. It only covers binary classification and two different regression methods, which require comparatively low processing power. Complicated methods such as neural nets and deep learning are out of the scope of the thesis, which may potentially improve observed results. Although comprehensive, the thesis is still not an end-to-end pipeline. It covers the modelling stage of the knowledge discovery cycle and only at a match or season level. Future research needs to apply techniques outlined by the current thesis on an event or play-by-play

data. Furthermore, The steps of pre-processing and visualizations need to be the focus of future research in conjugation with the findings of the current work. In conclusion, sports research needs to leverage big data for finding novel solutions to a wide array of problems across sports domains, and ML/DS methods seem to be the ideal tool for this. Specifically, the crucial steps required are the normalization of notational data, using multiple models for robustness and k-fold cross-validation for determining the out-of-sample validity of the findings. Furthermore, the thesis provides an introduction to how data science techniques and multidisciplinary approaches can help the sports industry and research.

Abstract: German

Big Data hat sich als zunehmend einflussreich in einer Vielzahl von Entscheidungsfindungs- und Erkenntnisgewinnung in verschiedenen Bereichen erwiesen. Die Sammlung, Vorverarbeitung und Analyse von Daten zur Entscheidungsfindung oder im Erkenntnisgewinnung ist eine komplexe Aufgabe. Mit dem zunehmenden Zugang zu Rechenleistung sind eine Vielzahl von Datenwissenschaftstechniken als Problemlösungswerkzeuge in vielen Bereichen der Gesellschaft entstanden. In den letzten zehn Jahren wurde eine große Menge an Daten im Bereich des Sports gesammelt. Mehrere Labore weltweit haben diese Daten in Verbindung mit Machine Learning/Data Science (ML/DS)-Techniken genutzt, um einen erheblichen Mehrwert für die Sportindustrie und die Wissenschaft zu schaffen. Im Vergleich zum Potenzial der verfügbaren Daten wurde jedoch nur ein kleiner Teil genutzt. Dies liegt hauptsächlich an mangelnden Kenntnissen in der Programmierung, die erforderlich sind, um die Daten in einer Form zu erhalten, die optimal ist, um Modelle zur Beantwortung spezifischer Fragestellungen von Interesse zu erstellen. Das Problem des Engpasses bei der Vorverarbeitung wurde teilweise durch Datenressourcen wie OPTA, STATSBomb und FBRef.com gelöst, die saubere Tracking-, Ereignis- und Notationsdaten bereitstellen. Darüber hinaus bieten Bibliotheken in Python und R wie Floodlight, AMIE und SoccerAction Pakete an, die Vorverarbeitungs- und Visualisierungsschritte vereinfachen und damit Experten in bestimmten Bereichen mit begrenzten Codierkenntnissen einen großen Zugang zur Big Data-Analyse bieten. Vor dem Hintergrund dieser Entwicklungen zielt die vorliegende Arbeit darauf ab, ML/DS-Methoden wie Regression, binäre Klassifikation, Feature-Engineering und k-fold Kreuzvalidierung in den Bereich der Sportanalytik einzuführen. Dies kann Fachleuten in bestimmten Bereichen die notwendigen technischen Werkzeuge bieten, um die steigende Datenmenge in der Sportindustrie zu nutzen.

Durch veröffentlichte Fallstudien, die jeweils eine spezifische Hypothese behandeln, erklärt die Arbeit die Bedeutung der Normalisierung von KPIs als Schritt des Feature-Engineering vor statistischer Modellierung. Sie erläutert auch den Wert der Verwendung von k-fold Kreuzvalidierung als Modellbewertungskriterium für Regression und Klassifikationsprobleme. Die Arbeit betont weiterhin den Wert der Verwendung mehrerer ML-Modelle zur Lösung spezifischer Probleme als Modellrobustheit, um falsche Ergebnisse aufgrund von Verzerrungen eines einzelnen Algorithmus zu vermeiden. Die bereitgestellten Methoden können potenziell auf die Forschung im Allgemeinen angewendet werden, aber der Bereich der Schlag- und Ballsportarten wie Cricket und Baseball scheint für die Big-Data-Analyse unter Verwendung von ML besonders geeignet zu sein. Dies liegt an ihrer einzigartigen geschlossenen Handlungsstruktur (eine Handlung, eine Reaktion, die zu einem Ergebnis dieses Handlungs-Reaktions-Paares führt), da sie im Vergleich zu Invasionssportarten einen geringeren Grad an Zufälligkeit aufweisen. Die vorliegende Arbeit hat jedoch einige Einschränkungen aufgrund ihres Umfangs. Sie umfasst nur die binäre Klassifikation und zwei verschiedene Regressionsmethoden, die vergleichsweise geringe Rechen-

leistung erfordern. Komplizierte Methoden wie neuronale Netze und Deep Learning fallen außerhalb des Rahmens der Arbeit, die möglicherweise zu verbesserten Ergebnissen führen könnten. Obwohl umfassend, stellt die Arbeit immer noch keine End-to-End-Pipeline dar. Sie behandelt lediglich die Modellierungsphase des Wissensentdeckungszyklus auf Match- oder Saisonebene. Zukünftige Forschung muss die in der vorliegenden Arbeit beschriebenen Techniken auf Ereignis- oder Spiel-für-Spiel-Daten anwenden. Darüber hinaus sollten die Schritte der Vorverarbeitung und Visualisierungen im Zusammenspiel mit den Ergebnissen der aktuellen Arbeit Gegenstand zukünftiger Forschung sein. Zusammenfassend muss die Sportforschung die Möglichkeiten von Big Data nutzen, um neuartige Lösungen für eine Vielzahl von Problemen in verschiedenen Sportbereichen zu finden, wobei ML/DS-Methoden das ideale Werkzeug dafür zu sein scheinen. Insbesondere sind die Normalisierung von Notationsdaten, die Verwendung mehrerer Modelle zur Robustheit und die k-fold-Kreuzvalidierung zur Bestimmung der außerhalb des Musters liegenden Gültigkeit der Ergebnisse wichtige Schritte. Darüber hinaus gibt die Arbeit eine Einführung in die Möglichkeiten, wie Datenwissenschaftstechniken und multidisziplinäre Ansätze der Sportindustrie und -forschung helfen können.

Overview of the Articles

Table 1 below provides an overview of the scientific publications the author has been involved in for the period of the PhD studies. A total of eight articles, of which five are first as author or first co-authors and three articles as co-author. All the mentioned articles have been published. The latest impact factors (IF) and quartiles of the journals have been mentioned wherever possible.

Table 1: List of published articles by the author of the thesis

Reference
Included articles from the author, on which the Introduction is based
Phatak, A.A., Wieland, FG., Vempala, K. et al. Artificial Intelligence Based Body Sensor Network Framework—Narrative Review: Proposing an End-to-End Framework using Wearable Sensors, Real-Time Location Systems and Artificial Intelligence/Machine Learning Algorithms for Data Collection, Data Mining and Knowledge Discovery in Sports and Healthcare. Sports Med - Open 7, 79 (2021). doi: 10.1186/s40798-021-00372-0. [6.766,Q1]
Included articles from the author, which are a part of the methods
Phatak, A.A., Mehta, S., Wieland, FG. et al. Context is key: normalization as a novel approach to sport specific preprocessing of KPI's for match analysis in soccer. Sci Rep 12, 1117 (2022). doi: 10.1038/s41598-022-05089-y. [4.996, Q1]
Phatak, A., Rein,R. & Memmert,D.(2021).The Dirty League: English Premier League Provides Higher Incentives for Fouling as Compared to other European Soccer Leagues. Journal of Human Kinetics,80(1) 263-276. doi: 10.2478/hukin-2021-0095. [2.923,Q1]
Jamil, M., Liu, H., Phatak, A., & Memmert, D. (2021). An investigation identifying which key performance indicators influence the chances of promotion to the elite leagues in professional European football. International Journal of Performance Analysis in Sport, 21(4), 641-650. doi: 10.1080/24748668.2021.1933845. [2.699, Q1]
Jamil, M., Phatak, A., Mehta, S. et al. Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. Sci Rep 11, 22703 (2021). doi: 10.1038/s41598-021-01187-5. [4.996, Q1]
Included articles from the author, which are cited in the discussion
Phatak, A., Mujumdar, U., Rein, R. et al. Better with each throw—a study on calibration and warm-up decrement of real-time consecutive basketball free throws in elite NBA athletes. Ger J Exerc Sport Res 50, 273–279 (2020). doi: 10.1007/s12662-020-00646-x. [1.086, Q2]
Jamil, M., Harkness, A., Mehta, S., Phatak, A., Memmert, D., & Beato, M. (2021). Investigating the impact age has on within-over and death bowling performances in international level 50-over cricket. Research in Sports Medicine, 1-10. doi: 10.1080/15438627.2021.1954515. [4.567, Q1]
Mehta, S., Phatak, A., Memmert, D., Kerruish, S., & Jamil, M. (2022). Seam or swing? Identifying the most effective type of bowling variation for fast bowlers in men's international 50-over cricket. Journal of Sports Sciences, 1-5. doi: 10.1080/02640414.2022.2094140 [3.943, Q2]

Contents

List of Figures

List of Tables

1 Introduction

1.1 Big data and the future

'Dataism' is a term popularized by Yuval Harari in the popular science book 'Homo Deus'(Harari, 2016). The term suggests that a wide array of decisions in the future pertaining to society will be taken through the analysis of 'Big Data. Big Data is currently defined as high-volume, high-velocity, high-variety and high-veracity information(4Vs). Where, volume refers to the magnitude or size of the data, variety refers to structural heterogeneity in the data set, velocity refers to the rate at which data are generated and veracity refers to the reliability of the data (Claudino et al., 2021; Rajšp & Fister, 2020; Roy et al., 2020). Big data as it stands holds tremendous untapped potential for use and applications in multiple industries such as health care, banking and finance, security, aviation, astronomy, agriculture, and sports (Rajšp & Fister, 2020; MacLennan, 2005; Rein & Memmert, 2016). Although big data can be an asset for the knowledge discovery process, using this data is a non-trivial task. Due to its unique nature in terms of the 4Vs' there arise challenges all the way from the collection of the data to deployment for a particular application. Noise accumulation, spurious correlation, measurement errors, and high computational power requirements are some of these challenges(Sagiroglu & Sinanc, 2013; Raghupathi et al., 2010). High-level conceptual end-to-end frameworks have been proposed to outline the complete cycle for knowledge discovery. The author of the current thesis has proposed one such framework, the Artificial Intelligence-Based Body Sensor Network Framework (AIBSNF) which describes the complete process from data collection to knowledge discovery for specific problems in the field of sports analytics (A. A. Phatak, Wieland, Vempala, Volkmar, & Memmert, 2021). The current thesis will only focus on the aspects post-acquisition of data viz. normalization, pre-processing and modelling using multiple machine learning algorithms for knowledge discovery, specifically in the sport of soccer.

1.2 Current status of data analytics in sports

The use of big data and AI/ML tools in sports were beach-headed in track and field, and weightlifting (Taborri et al., 2020). The first sport to use data for recruiting and performance-enhancing purposes were Baseball(MacLennan, 2005). Basketball and football soon caught up, as several professional teams and academics started using 'big data for recruiting, performance analysis, and performance enhancement(Rico-González, Pino-Ortega, Méndez, Clemente, & Baca, 2022). The collected data in the field of sports had a high variability due to the wide range of sports it has been used in. This data can be categorised as physiological data, position tracking data, psychological data, scouting data and video data(Rein & Memmert, 2016). Table.2 below shows a non-exhaustive list of data sources for the sport of soccer, either in the form of companies which provide data

and related services or open-source websites where data can be obtained under a creative commons licence. The companies listed below provide a wide set of services ranging from raw data to analysis and visualization. This has been done across multiple sports-related fields such as betting, self and opponent analysis, coaching, player load management, recruiting etc. The availability of such data has greatly increased the potential for research in sports analytics. Due to this, there seems to be a need to outline methods to effectively use the data for the process of knowledge discovery across sports and domains(Rein & Memmert, 2016; Bai & Bai, 2021). A subset of data from the sources listed below has been used in published studies included in the current thesis.

Table 2: Non-exhaustive list of open data sources for Soccer

Company name	**Reference**	**Data type**
data providing companies: primary sources		
OPTA Sports	https://www.optasports.com/	In-game Event & Tracking Data
Hudl	https://www.hudl.com/	Video data & analysis
Instat	https://instatsport.com/	Event & Tracking Data
Statsbomb	https://statsbomb.com/	Event & Tracking Data
Stats Perform	https://www.statsperform.com/	Event & Tracking Data
Sportradar	https://www.sportradar.com/	End to end services
Wyscout	https://wyscout.com/	Event & Tracking Data
Kitmanlabs	https://www.kitmanlabs.com/	Player tracking and load management
Catapult Sports	https://www.catapultsports.com/	End to End Services
Open source data websites		
Whoscored	https://whoscored.com/	In game statistics for Soccer
FBref	https://fbref.com/	In game statistics for Soccer
Scientific Literature offering open source data		
Published Data set	(Pappalardo et al., 2019)	Match Event Data
Published Data set	(Dubitzky, Lopes, Davis, & Berrar, 2019)	Game results data
Published Data set	(Biermann et al., 2021)	Event Data

Sports data currently is collected and stored primarily in a few different formats. Game video data, spatial-temporal position tracking data, biometric data, event data, coaching data, scouting data and psychological data. This data can be further cleaned and filtered into specific data sets data on the game or season level (notational data). This wide array of data exists across the spectrum from being highly structured to unstructured(Rein & Memmert, 2016). For the scope of the thesis, we primarily focus on notational data. Currently, the primary method for collection of tracking is Real-Time Location System (RTLS) and Computer vision. Manual event tagging is performed through video analysis to get notational data as a post-processing step. Different methods have different chal-

lenges such as tracking in crowded situations, lack of physiological data, difficulty tracking high-speed objects and also human error depending on the tracking technology used for data collection (Vijayakumar & Nedunchezhian, 2012; Bialkowski et al., 2014).

Furthermore, the sheer volume of such data makes it difficult for mining relevant information from the noise(Bialkowski et al., 2014). Further improvements in collecting spatial-temporal data such as computer vision, wearable sensors and RTLS can potentially give access to large volumes of data in real-time. Artificial Intelligence and Machine Learning (AI/ML) seem to be effective tools which have the potential to not only address the challenges but also extract knowledge from data arising from the improvements in mentioned technologies(A. A. Phatak, Wieland, et al., 2021). Hence, insights into Data Science (DS) methods may prove useful to further the knowledge discovery process in a multitude of sports. The current thesis only uses in-game statistics (notational data) obtained directly or indirectly from OPTA sports which have been scientifically validated for having high accuracy and reliability(Jamil, 2019; Liu, Hopkins, Gómez, & Molinuevo, 2013; Jamil, Liu, Phatak, & Memmert, 2021).

1.3 Machine learning and sports analytics

Over the past two decades, improvements in data collection technology, ML methods, increasing computational power and availability of tools such as inbuilt libraries in python and R there have boosted the use of ML/DS methods in sports. This wave has hit the professional sports industry and academics alike (sports science)(A. A. Phatak, Wieland, et al., 2021; Swartz, 2020; Rein & Memmert, 2016; Hao & Ho, 2019). There has been a rise in conferences such as MIT SLOAN purely dedicated to data analysis in sports, with a large subset of the publications based on applications of AI/ML methods in sports(Sloan, 2017). Furthermore, several articles have been published across several multidisciplinary journals exploring data analytics in sports (Swartz, 2020). A PubMed search using 'sports analytics' shows a total of 4348 hits as of 3.8.2022, with 3539 of them published post 2010. A recent publication which reviewed machine learning applications specifically in soccer summarised 32 studies which investigated injury prediction, performance and talent forecasting using ML algorithms. It recommended exploring the required size of data to make relevant and accurate predictions across the above-mentioned fields. Table 3 below outlines some of these studies which use ML methods in the domain of soccer tactics, which is the primary area of interest of the current thesis.

Table 3: A non-exhaustive list of publications which used Machine Learning for performance analysis in soccer

Topic	**Central concept**	**Reference**

VAEP & XT	Ranking threat of teams and players based on their on the ball actions, based on if they increase or decrease the chances of goal scoring using gradient boosting classification model	(Decroos, Bransen, Van Haaren, & Davis, 2020; Van Roy, Robberechts, Decroos, & Davis, 2020; Decroos, Bransen, Van Haaren, & Davis, 2019)
Which pass is better	Assess effectiveness of a pass by evaluating their impact on majority situations and space control in front of the goal using Voronoi-diagrams.	(Rein, Raabe, & Memmert, 2017)
Not all passes are created equal	Estimating risk and reward of a pass on a continuous scale using tracking data and outlining multiple used cases for evaluating player performances	(Power, Ruiz, Wei, & Lucey, 2017)
Incorporating domain knowledge in ML models for outcome prediction	Feature extraction and learning by incorporating domain-specific knowledge to improve model performance	(Berrar, Lopes, & Dubitzky, 2019)
DCNN to predict Goal-Scoring Opportunities in Soccer	Using GoogLeNet and 3 layered CNN architecture to classify enhanced images of ball possession phases to find which one of them ends up on a shot on goal.	(Wagenaar, Okafor, Frencken, & Wiering, 2017)
Predicting match outcome according to the quality of the opponent in the English premier league using situational variables and team performance indicators	Decision trees performed the best while predicting the result of matches when an input feature of opponent team strength was provided along with 22 situational variables and performance indicators.	(Bilek & Ulas, 2019)
Quantifying the relation between performance and success in soccer	Using ML algorithms and technical data the study simulated outcomes of an entire season based on 6396 games and 10 million events from top 6 European soccer leagues	(Pappalardo & Cintia, 2018)
Classification of Passes in Football Matches using Spatio-temporal Data	Classification model created with features extracted from Spatio-temporal data using computational geometry was able to classify passes as 'Bad', 'Ok' and 'Good' with and accuracy of 85.8%.	(?, ?)
Not Every Pass Can Be an Assist	Quantifying the effectiveness of passes based on measures evaluated from tracking data which can be used for analyzing complex dynamics in build-up play and space creation in soccer.	(Goes, Kempe, Meerhoff, & Lemmink, 2019)

Discovering Team Structures in Soccer from Spatio-temporal Data	Classified teams and ranked players based on heat maps and pass locations with an accuracy of 87%. The study also predicted occurrence shots for each possession phase with AUCROC of 0.785	(Brooks, Kerr, & Guttag, 2016)
Using ML to draw inferences from pass location data in soccer	Applying unsupervised learning on 21.5 million frames of player tracking data the study presents a method for automatic formation template (team structure) detection.	(Bialkowski et al., 2016)
Learning to Rate Player Positioning in Soccer	The study describes a purely data driven approach using deep reinforcement learning to rate multiplayer positions based on tracking data. It validates this by comparing the automated ratings with actual dangerous situation in the game.	(Dick & Brefeld, 2019)
Spatio-temporal convolution kernels	The study propose a novel class of Spatio-temporal convolution kernels which capture similarities in multi-object scenarios. Furthermore, it compares it to baseline techniques for clustering real and simulated team sport data.	(Knauf, Memmert, & Brefeld, 2016)
Finding efficient strategies in 3-versus-2 small-sided games of youth soccer players	The study used unsupervised learning to find efficiency of tactical patterns using 30 technical and tactical parameters for a shot and a event right before the shot occurred.	(Leser et al., 2019)
Individual ball possession in soccer	The study automatically classified ball possession phases into six different categories using Bayesian Network based on features pertaining to the position and mechanics of the ball and the players. The F-score for all phases were between 0.77 and 0.88	(Link & Hoernig, 2017)
Team activity recognition in soccer	Automatic recognition from match video data was performed using the bag of Words technique. NN-classifier, KNN, SVM and RF were used to classify video clips as Ball Possession, Quick Attack and Set Piece. The RF performed the best in one of the experiments with an accuracy of 92.89 %.	(Montoliu, Martín-Félez, Torres-Sospedra, & Martínez-Usó, 2015)

ML and Wearable Sensors to Predict Energetics and Kinematics	Four supervised ML techniques were used to predict turn direction, speed, and mechanical work based on 18 features derived from IMU data. The classification of turning direction yielded a score of >98.4% while the regression task showed R^2 of 0.42-0.43 for predicting mechanical work and 0.66-0.69 for running speed before and after the turn.	(Zago, Sforza, Dolci, Tarabini, & Galli, 2019)
The tactics of successful attacks in professional Soccer	The study analyzes the behaviour of dynamic subgroups in relation to successful attacks. The Unsupervised classification was used to classify attacks as successful or unsuccessful. Successful attacks strongly depend on defenders creating space for the attackers along with their synchronization with midfielders.	(Goes, Brink, Elferink-Gemser, Kempe, & Lemmink, 2021)

1.4 The bottleneck

A review, published in 2013, proposed a framework based on a review of contemporary data mining techniques in elite sports(Ofoghi, Zeleznikow, MacMahon, & Raab, 2013). It aimed at reviewing and understanding, the common data mining techniques relevant to elite sports analytics such as clustering, classification, relationship modelling, regression analysis, and rule mining. The same study outlined specific requirements of pre-processing steps and model evaluation methods. It outlined a need for the application of data mining methods in the field of sports performance analysis(Ofoghi et al., 2013). Furthermore, a high amount of pre-processing and data visualization steps are required to prepare the data in a form where relevant and interesting research can be conducted and presented. The technical coding skills required for these steps are generally possessed by computer scientists, who seem to lack the domain-specific expertise to formulate relevant questions in the field of sports(Rein & Memmert, 2016).

Multidisciplinary knowledge in the fields of computer science and sports science seems to be necessary to conduct research analogous to those mentioned in table 3 above. Although considerable research has been done using DS and ML methods, the expertise and computation power and computer science skills required to perform such research are still high. Currently, most sports science institutes lack this(Rein & Memmert, 2016). In the past decade, the availability of clean high quality data has increased thanks to sources mentioned in table 2. Open source python packages such as floodlight (Power et al., 2017), mlp soccer (https://pypi.org/project/mplsoccer/), SoccerAction(?, ?), SoccerMix(Decroos, Roy, & Davis, 2020), AMIE(Decroos, Schütte, Beéck, Vanwanseele,

& Davis, 2018) (https://dtai.cs.kuleuven.be/sports/software) have streamlined the number of data cleaning and visualization steps (Goes et al., 2019). Out of the box, DS/ML techniques can be now applied for solving a wide array of research questions with relative ease thanks to libraries such as Scikit-learn (python) and Caret (R) (Kuhn, 2008; ?, ?). These developments have opened a whole array of data science methods as potential applications in the field of sports and has partly solved the problem of the pre-processing bottleneck.

1.5 Big data and statistical robustness

The standard practice in research in general is to use frequentist statistical methods such as t-tests or ANOVA, which have been ideally built for cross-over designed studies (Hecksteden, Faude, Meyer, & Donath, 2018). Although, this design is the gold standard for sports research, in reality, a very low number of studies follow this due to the practical issues with participants and the time investment needed. In such situations, it is common to have small sample sizes. In these cases, frequentist statistics and p-value cut-offs of 0.05 can be justified but, there are several other issues with this structure. Lack of replication of research, publication bias towards positive results, under-powered studies, falsely interpreting results, exaggerating or dismissing genuine effects are some of them (Hecksteden et al., 2018; Amrhein, Greenland, & McShane, 2019).

With the rise of big data in research, experiments involving a very high number of data points(form $> 10^3$) have become possible. A large amount of data is generally considered an advantage in research due to its high power, but this also has a number of disadvantages that need to be accounted for. Big data sets have a tendency of showing random correlations due to the sheer volume of data. Hence, the standard p-value cut-off of 0.05 needs to be reevaluated based on the number of data points used in the study (Lin, Lucas Jr, & Shmueli, 2013). Another issue is multicollinearity within the dependent variables, this can be an issue for traditional statistical approaches as this tends to inflate variance and coefficients in regression models. Most of these issues can be potentially dealt with by using big data to account for power and cross-validation to investigate the out-of-sample validity of the chosen model(s) and multiple model approaches to avoid bias of a single model. (de Rooij & Weeda, 2020).

1.6 Objectives of the current research

The current thesis as a whole aims to introduce DS and Supervised Machine Learning methods into sports research for domain experts. The research primarily focuses on the sport of soccer and investigates KPIs using in-game statistics to explain the need for robustness testing while handling large data sets. It applies simple methods such as scaling, feature extraction, classification and, regression with special emphasis on normalizing and cross-validation using published case studies. The thesis puts an emphasis on how these

techniques could be applied for knowledge discovery across different sports. Figure 1 below shows the scope of the methods covered in the DS pipeline study based on knowledge discovery. The Specific aims/hypothesise of each study will be investigated in each of the respective studies as mentioned in table 4 below. The emphasis of the thesis is not so much on the specific aims, but rather on the impact of the methods on current and future research (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Figure 1: Knowledge discovery process and scope of the study (highlighted in green)

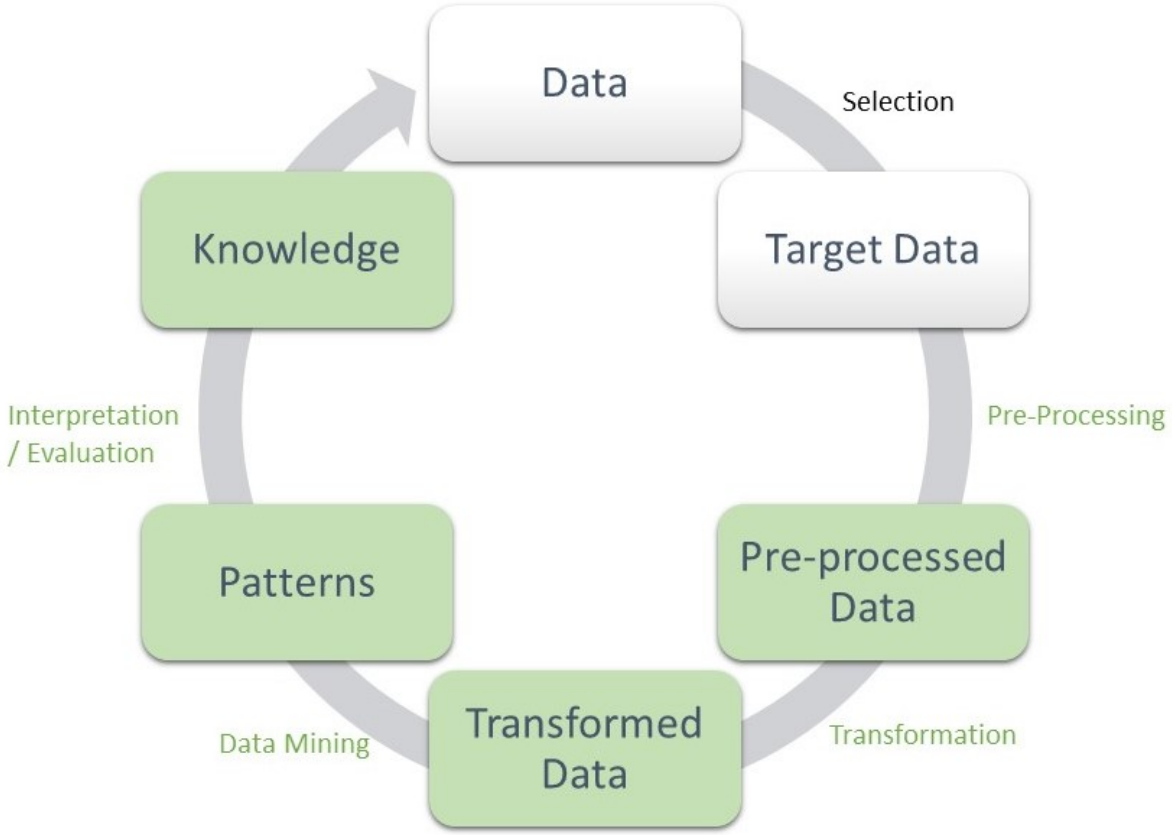


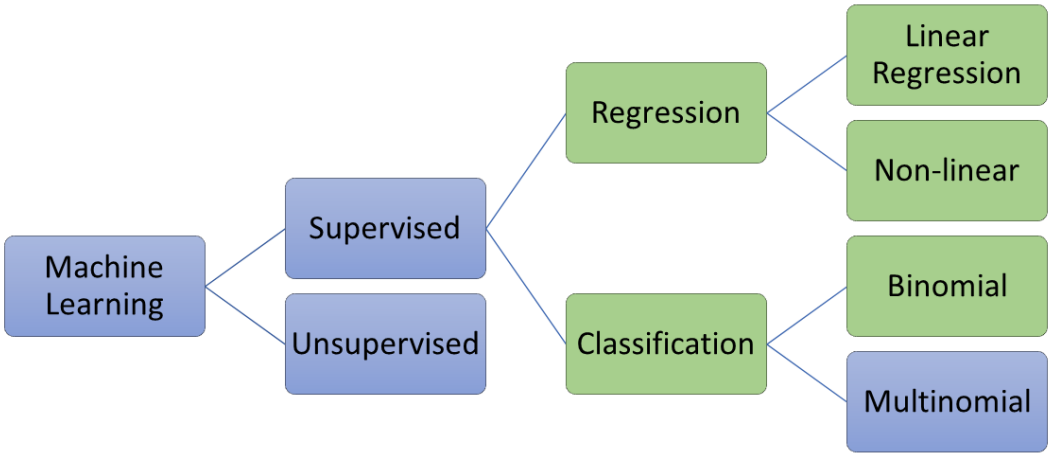
Table 4: List of Aims the studies included in the thesis have investigated

Index	Aim/Hypothesis	Reference
Pre-processing		
A1	The aim of the 'Context is key' study is to outline the importance of normalizing KPI's according to sport specific context for soccer and to provide for the first-time data pre-processing technique for effectively incorporating sport specific context within the sport(A. A. Phatak et al., 2022).	Case study I (see methods)
Regression		
H1	(H1) The Dirty League study predicts a positive correlation of both Normalised Fouls per game (FPGNorm) and Yellow card fouls per game (YCFNorm) with the end of season points (Pts) and a negative correlation with the number of goals conceded (GA)(A. A. Phatak, Rein, & Memmert, 2021).	Case study II (see methods)
H2	The effects of fouling on performance will be significantly different in the EPL as compared to the other leagues(A. A. Phatak, Rein, & Memmert, 2021).	Case study II (see methods)
H3	The ratio of yellow cards per foul per game (YCPFPG) will show a negative correlation with the end of season points and positive correlation with GA, with the EPL showing significantly different effects as compared to the other leagues(A. A. Phatak, Rein, & Memmert, 2021).	Case study II (see methods)
Classification		
A3	The promotion relegation study aims to utilise predictive statistical models to assess match-level technical performance data for 98 teams performing in the English Championship, French Ligue 2 and German Bundesliga 2 over a five season sample period (2013/14 to 2017/18), to identify the technical KPI's most likely to aid the chances of promotion to the elite leagues (Jamil, Liu, et al., 2021).	Case study III (see methods)
A2	The aim of the Goalkeeping study is to present a multiple model approach to classify elite goalkeepers from performance data and identify features, which distinguish them from their sub-elite counterparts(Jamil, Phatak, et al., 2021).	Case study IV (see methods)

2 Methods & case studies

ML algorithms can be primarily classified as Supervised (SL) and Unsupervised (US). For the scope of the current thesis, we focus on SL algorithms. These primarily include linear regression, gradient boosting machine, logistic regression and random forest primarily for regression and binary classification tasks. The overall methodological scope of the thesis has been highlighted in green in fig 2 below.

Figure 2: Methodological Scope of the thesis from an ML perspective (highlighted in green)



2.1 Data

Primarily, data from OPTA Sports has been used in conjunction with open-source data from FBref.com and Whoscored.com. The selection criteria and data subsets used for each of the studies has been described in the respective articles.

2.2 Domain specific normalization

Data collected from the real world is not perfect. It may be inconsistent, noisy and have missing values. Pre-processing methods such as cleaning, integration, transformation and, reduction can be used for getting data into a usable format for relevant analysis(Alasadi & Bhaya, 2017). In addition to generalized issues, each domain has its own challenges. In the domain of sports, evening out of data and using raw data can lead to issues such as reliability, validity and, errors in the precision of the results(Ali, 2011; O'Donoghue,

2007). In order to build the context from the domain into the problem at hand, a wide range of domain-specific pre-processing steps can be used(A. A. Phatak et al., 2022). One such method has been outlined below, which can be used for performance analysis in soccer. It controls for team strength by normalizing defensive KPIs with possession as the normalizing variable. Equation 1 below outlines the central concept of the study, which is fully elaborated in case study 1 along with the analysis and results of the A1 (see Table 4.).

$$NormDefensiveKPI = \frac{DefensiveKPI}{(1 - PossessionPercentage/100)} \tag{1}$$

2.2.1 Case study I

Phatak, A.A., Mehta, S., Wieland, FG. et al. Context is key: normalization as a novel approach to sport specific preprocessing of KPI's for match analysis in soccer. Sci Rep 12, 1117 (2022). https://doi.org/10.1038/s41598-022-05089-y

2.3 Linear Regression

In the second case study, the normalization technique elaborated above is deployed to investigate incentives for out-of-possession fouling in the top 5 European leagues. This is modelled using multiple linear regression analysis with interaction effects. The study emphasizes the value of normalizing fouls and yellow cards for out-of-possession time on a season level. These normalized features are then fed into three linear regression models in order to investigate the incentives of fouling in the English Premier League as compared to other top European soccer leagues. Case study II below elaborates on this concept further.

2.3.1 Case study II

Phatak,A.,Rein,R. Memmert,D.(2021).The Dirty League: English Premier League Provides Higher Incentives for Fouling as Compared to other European Soccer Leagues. Journal of Human Kinetics,80(1) 263-276. https://doi.org/10.2478/hukin-2021-0095

2.4 Logistic Regression

Case study III below uses binary Logistic regression to find which KPIs are most relevant in determining the promotion of teams from second to first division in elite men's European soccer leagues.

2.4.1 Case study III

Jamil, M., Liu, H., Phatak, A., Memmert, D. (2021). An investigation identifying which key performance indicators influence the chances of promotion to the elite leagues in

professional European football. International Journal of Performance Analysis in Sport, 21(4), 641-650. https://doi.org/10.1080/24748668.2021.1933845

2.5 Binary classification: Multiple ML classifiers

Three different binary classification algorithms were used to distinguish elite goalkeepers from their sub-elite counterparts in professional men's soccer. The study outlines a multiple-model approach not only to classify the goalkeepers, but also to provide insight into which KPIs were the most important as a distinguishing criterion. Case study IV below elaborates on this further.

2.5.1 Case study IV

Jamil, M., Phatak, A., Mehta, S. et al. Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. Sci Rep 11, 22703 (2021). https://doi.org/10.1038/s41598-021-01187-5

3 Discussion

The discussion of the specific aims and hypotheses from the thesis have been addressed in the case studies within the methods section. The sections below elaborate on the implications of the thesis as a whole in regrades to its overall contribution to sports performance analysis and scientific practice in general.

3.1 Value of normalizing

Feature engineering is a crucial step in the knowledge discovery pipeline and is often overlooked within the sports domain(Zheng & Casari, 2018; Duboue, 2020). The choice of variables used as input features affects the performance of the model in multiple ways. One important point of contention is the trade-off between interpretability and transparency while choosing the complexity of the model(Lipton, 2018). The domain of sports is no exception to this phenomenon (Sun, Davis, Schulte, & Liu, 2020). Performance analysis across team sports involves analyzing performance indices. This may be on a player level, a team level, a play-by-play level or on a season level. In every case, each performance index used as an input feature (independent variable) needs to be taken within its respective context. Case studies I and II in the methods section elaborate on the specific normalization method used, for analyzing defensive KPIs, and their impact on success. The normalization was performed on a season level, using possession as a proxy for team strength (A. A. Phatak et al., 2022). This phenomenon is of great relevance in subdomains within sports such as self/opponent analysis, recruiting, scouting and youth development. One of the challenges in all of these areas is that the teams are of different strengths. Hence, controlling for team strength while comparing players or specific KPIs becomes crucial (A. A. Phatak et al., 2022). The same idea can be used to perform analysis on a more granular level (play-by-play or match-by-match). Furthermore, team strength proxies such as betting odds, table position and points can be used as normalizing variables(Klemp, Wunderlich, & Memmert, 2021). On a player level, the same concept can be used from measuring efficiency in different phases of the game. One such example for measuring attacking efficiency would be the number of assists or key passes normalized per 'ball touch' of the player in question.

Overall, building in context with respect to the problem in hand is a crucial concept. So far, this has not been widely implemented while conducting performance analysis in sports using big data. The idea of normalization used in the study falls under the category of (domain-specific) feature engineering/feature transformation from a data science perspective. Although transforming features to obtain interpretive models is a common practice in data science, it has not yet been used in sports analytics to a wide extent. Furthermore, the mathematical transformation of features to improve model performance is another common feature engineering practice, which is still unpopular in sports ana-

lytics. One example of this idea is the transformation of Cartesian coordinates into polar to model expected goals in soccer(Robberechts & Davis, 2020). There seems to be a need to propagate the idea of feature engineering using domain-specific knowledge and mathematical transformations to implicitly incorporate domain specific knowledge into the field of sports analytics.

3.2 Multiple machine learning algorithm approach

Sports statistics encode a small subset of information within specific sports. Most invasion sports are dynamic in nature, and thus the data creation process (sport itself) can potentially be linear or non-linear(Paul, Bradley, & Nassis, 2015). This may result in the features of the data set being multicollinear to each other. Traditional frequentist statistics may not necessarily be able to handle this, but algorithms such as random forest and gradient boosting machines can handle multicollinearity (within the input features) and generate non-linear models(Razavi, Gill, Åhlfeldt, & Shahsavar, 2005). Using a single such model may not suffice, as MLAs are usually optimized through stochastic approaches(Oza & Tumer, 2008). This may result in different algorithms giving different results even when used on the same data set to solve the same problem. As shown in previous research, multiple model approaches can successfully offer robust methods against bias-variance trade-off (Schapire & Singer, 1999). Multiple model approaches have been so far used across industries but, there seems to be a lack of applications in the area of sports research. Thus, it may be possible to improve the consistency of results within the scientific practice by incorporating a multiple model approach.

3.3 K-fold cross-validation for out-of-sample validity

Case studies 1, 2 and 4 use k-fold cross-validation for model evaluation. The mean and standard deviation for The mean of the selected evaluation matrices were reported to depict the out-of-sample validity of the trained models. K-fold cross-validation (CV) is a method where the data is split into k equal parts. Each of the k-sets is used for as testing, while the other k-1 sets are used for training the model. The whole procedure is repeated k times for optimization matrices of choice (accuracy, f1, R^2, specificity) for each round (Mosteller & Tukey, 1968). CV can be applied in the context of model evaluation, model selection and tuning hyperparameters of selected algorithms (Refaeilzadeh, Tang, & Liu, 2009). It can also be used as an alternative to hypothesis testing for checking the out-of-sample validity of the chosen model. The advantage of using CV is that it does not rely on any assumptions, as opposed to hypothesis testing (de Rooij & Weeda, 2020). CV is model agnostic due to the fact that the splitting of the data set into k-folds is performed before the model is trained. Hence, any prediction model regardless of its complexity can theoretically be cross-validated to check it's out-of-sample validity (Mosteller & Tukey, 1968). Furthermore, algorithms such as grid search and random search CV could also be

used to tune the hyperparameters of almost all ML algorithms, resulting in better models for specific data sets. (Refaeilzadeh et al., 2009).

There has been considerable debate in retiring hypothesis testing. Statistical significance and 'real world effect' have been misconstrued, and this may lead to false conclusions. Using the terms 'statistically significant' or 'non-significant' may imply effects where none exist and vice versa(Amrhein et al., 2019). With regard to big data, there is a high chance of random significant correlations arising due to sheer data volume, this may lead to the p-value cutoff of 0.05 being too high (Wasserstein & Lazar, 2016). CV offers an explanation which can check out of sample validity of the used model. It also shows what rate of error is expected when extrapolating the model for out-of-sample data. The tolerance of error for each problem is different based on the possible risk of obtaining false results. CV offers a way of assessing this through the standard deviation scores of the chosen optimization parameter including but not limited to F1,R^2, and AUC, depending on the specific requirements. These properties make CV an intuitive and transparent method for assessing any model, as compared to contemporary hypothesis testing through p-value cutoffs. Interpreting CV results to draw valuable insights from the models is a non-trivial task and requires a deep understanding of the data and the specific problem. Nonetheless, reporting CV scores is a valuable addition to standard scientific practice in big data analysis(de Rooij & Weeda, 2020).

3.4 Prospects in other team sports

Apart from soccer, multiple fields of sport have used big data and/or machine learning for investigating different aspects of the sport such as injury prediction, performance analysis, recruiting, betting, outcome prediction and investigation of motor control(?, ?; A. Phatak et al., 2020). Each sport has its own unique challenges when it comes to the nature of the data available its granularity, the pre-processing steps required for solving the chosen problem and the corresponding computational needs. The degree of randomness inherent in the sport is a crucial factor(Lopez, Matthews, & Baumer, 2018). It determines how much knowledge can be extracted from the given data and its theoretical limit(Severini, 2020).

Sports like cricket and baseball, categorized as bat and ball sports, provide an ideal environment for using match statistics for performance analysis. The nature of bat and ball games is such that the record of each pitch or delivery contains information about the overall game state. Performance of the batsman/hitter, performance of the bowler/pitcher and constellation of the team (outfield). Hence, the performance of the team and individual players can be reflected in the match statistics better than in invasion sports such as soccer, which are more dynamic as compared to bat and ball sports(Raja, Manasa, Reddy, & Sundari, 2021; ?, ?; Rein & Memmert, 2016). This seems to be conducive for data analytics, as the randomness in bat and ball sports is considerably lower than that in invasion

sports (Wunderlich, Seck, & Memmert, 2021). Furthermore, thanks to the data provided by companies such as OPTA sports, ball-by-ball analysis is made possible in cricket and baseball. So far, few studies have done this type of big data analysis in cricket. An of example such research is the study which investigates effects of age on bowling economy within 96 international elite bowlers(Jamil, Liu, et al., 2021). Another example of this is a study which examines 13176 balls bowled by international level fast blowers to investigate the relationship of delivery type and its effectiveness(Mehta, Phatak, Memmert, Kerruish, & Jamil, 2022). These studies are initial attempts at the use of 'big data' for research within bat and ball sports. There is a lot of room for using ML methods to investigate available data sets to gain new and valuable insights within bat and ball sports.

3.5 Limitations

In spite of having a number of pre-processing steps and algorithms covered in the thesis, the thesis does not build a complete end-to-end pipeline. Pre-processing steps and the choice of algorithms change from problem to problem. Hence, extrapolating the outlined techniques to other sports requires a considerable level of domain and data science knowledge. Another drawback of most of the included methods is that they work optimally on large data sets. The whole field of visualizing results has not been covered in the thesis, which is a crucial step in the communication of research findings to the stakeholders. This step needs to be appended to the findings of the current thesis in order to propagate the discussed ideas. Another point is that the data used in the current thesis is mostly aggregated data, either on a game or a season level, there is a need to do a similar analysis on data at play-by-play level(Klemp, Memmert, & Rein, 2022; A. A. Phatak et al., 2022). The domains of position data and video analysis are out of the scope of the current thesis, which forms a large part of performance analysis. Due to limited computation power, the area of deep learning, neural networks and other computation-intensive techniques have not been explored which potentially can improve the results of the included studies. Overall, due to the vastness of available DS methods and a large diversity withing data sets available the thesis only focused on the simplest solutions for chosen problems.

3.6 Conclusion

There seems to be tremendous potential in the multidisciplinary approach of using data science techniques for performance analysis and training science in general. Specifically, analyzing video data, event data, and spatial-temporal tracking data in soccer along with other invasion sports. The methods outlined in the current thesis such as normalization, CV and robustness testing using a multiple-model approach can prove highly valuable not just in the field of sports analytics but also research in general. There seems to be a need to build an end-to-end framework for conducting sports research on big data and communicating it to the stakeholders. The current thesis mainly addresses the statistical

modelling part of the framework. Future research should focus on streamlining pre-processing steps and building efficient infrastructure to visualize the results of the models as a supplement to the steps of the current thesis. The concepts outlined should also be investigated on more granular data sets and on other sports involving varying degrees of randomness. In essence, the pre-processing challenges and opportunities with the rise of big data in sports(Rein & Memmert, 2016). The current thesis offers a window into how simple methods from data science can greatly benefit knowledge discovery and decision-making processes across a wide range of sports in both research and its transfer in the industry.

A Appendix

A.1 Article I

Phatak, A.A., Wieland, FG., Vempala, K. et al. Artificial Intelligence Based Body Sensor Network Framework—Narrative Review: Proposing an End-to-End Framework using Wearable Sensors, Real-Time Location Systems and Artificial Intelligence/Machine Learning Algorithms for Data Collection, Data Mining and Knowledge Discovery in Sports and Healthcare. Sports Med - Open 7, 79 (2021). https://doi.org/10.1186/s40798-021-00372-0

A.2 Article II

Phatak, A., Mujumdar, U., Rein, R. et al. Better with each throw—a study on calibration and warm-up decrement of real-time consecutive basketball free throws in elite NBA athletes. Ger J Exerc Sport Res 50, 273–279 (2020). https://doi.org/10.1007/s12662-020-00646-x

A.3 Article III

Jamil, M., Harkness, A., Mehta, S., Phatak, A., Memmert, D., Beato, M. (2021). Investigating the impact age has on within-over and death bowling performances in international level 50-over cricket. Research in Sports Medicine, 1-10. https://doi.org/10.1080/15438627.2021.1954515

References

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102–4107. doi: 10.36478/jeasci.2017.4102.4107

Ali, A. (2011). Measuring soccer skill performance: a review. *Scandinavian Journal of Medicine & Science in Sports*, *21*(2), 170–183. doi: 10.1111/j.1600-0838.2010.01256.x

Amrhein, V., Greenland, S., & McShane, B. (2019). *Scientists rise up against statistical significance.* Nature Publishing Group. doi: 10.1038/d41586-019-00857-9

Bai, Z., & Bai, X. (2021). Sports big data: Management, analysis, applications, and challenges. *Complexity*, *2021*, 11. doi: 10.1155/2021/6676297

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine learning*, *108*(1), 97-126. doi: 10.1007/s10994-018-5747-8

Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S., & Fookes, C. (2016). Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), 2596–2605. doi: 10.1109/TKDE.2016.2581158

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., & Matthews, I. (2014). Large-scale analysis of soccer matches using spatiotemporal tracking data. In *2014 ieee international conference on data mining* (pp. 725–730). doi: 10.1109/ICDM.2014.133

Biermann, H., Theiner, J., Bassek, M., Raabe, D., Memmert, D., & Ewerth, R. (2021). A unified taxonomy and multimodal dataset for events in invasion games. In *Proceedings of the 4th international workshop on multimedia content analysis in sports* (p. 1–10). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3475722.3482792

Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the english premier league using situational variables and team performance indicators. *International Journal of Performance Analysis in Sport*, *19*(6), 930–941. doi: 10.1080/24748668.2019.1684773

Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(5), 338–349. doi: 10.1002/sam.11318

Claudino, J. G., Cardoso Filho, C. A., Boullosa, D., Lima-Alves, A., Carrion, G. R., GianonI, R. L. d. S., ... Serrão, J. C. (2021). The role of veracity on the load monitoring of professional soccer players: A systematic review in the face of the big data era. *Applied Sciences*, *11*(14). doi: 10.3390/app11146479

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (p. 1851–1861). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3292500.3330758

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2020). Vaep: an objective approach to valuing on-the-ball actions in soccer. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, ijcai-20* (pp. 4696–4700). doi: 10.24963/ijcai.2020/648

Decroos, T., Roy, M. V., & Davis, J. (2020). Soccermix: representing soccer actions with mixture models. In *Joint european conference on machine learning and knowledge discovery in databases* (Vol. 12461, pp. 459–474). doi: 10.1007/978-3-030-67670-4_28

Decroos, T., Schütte, K., Beéck, T. O. D., Vanwanseele, B., & Davis, J. (2018). Amie: Automatic monitoring of indoor exercises. In *Joint european conference on machine learning and knowledge discovery in databases* (Vol. 11053, pp. 424–439). doi: 10.1007/978-3-030-10997-4_26

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. doi: 10.1177/2515245919898466

Dick, U., & Brefeld, U. (2019). Learning to rate player positioning in soccer. *Big data*, *7*(1), 71–82. doi: 10.1089/big.2018.0054

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The open international soccer database for machine learning. *Machine Learning*, *108*(1), 9–28. doi: 10.1007/s10994-018-5726-0

Duboue, P. (2020). *The art of feature engineering: essentials for machine learning*. Cambridge University Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37. doi: 10.1609/aimag.v17i3.1230

Goes, F. R., Brink, M. S., Elferink-Gemser, M. T., Kempe, M., & Lemmink, K. A. (2021). The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, *39*(5), 523–532. doi: 10.1080/02640414.2020.1834689

Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. (2019). Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big data*, *7*(1), 57–70. doi: 10.1089/big.2018.0067

Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361. doi: 10.3102/1076998619832248

Harari, Y. N. (2016). *Homo deus: A brief history of tomorrow*. Random House.

Hecksteden, A., Faude, O., Meyer, T., & Donath, L. (2018). How to construct, conduct and analyze an exercise training study? *Frontiers in physiology*, *9*, 1007. doi: 10.3389/fphys.2018.01007

Jamil, M. (2019). A case study assessing possession regain patterns in english premier league football. *International Journal of Performance Analysis in Sport*, *19*(6), 1011–1025. doi: 10.1080/24748668.2019.1689752

Jamil, M., Liu, H., Phatak, A., & Memmert, D. (2021). An investigation identifying which key performance indicators influence the chances of promotion to the elite leagues in professional european football. *International Journal of Performance Analysis in Sport*, *21*(4), 641–650. doi: 10.1080/24748668.2021.1933845

Jamil, M., Phatak, A., Mehta, S., Beato, M., Memmert, D., & Connor, M. (2021). Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. *Scientific reports*, *11*(1), 1-7. doi: 10.1038/s41598-021-01187-5

Klemp, M., Memmert, D., & Rein, R. (2022). The influence of running performance on scoring the first goal in a soccer match. *International Journal of Sports Science & Coaching*, *17*(3), 558–567. doi: 10.1177/17479541211035382

Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, *11*(1), 1–10. doi: 10.1038/s41598-021-03157-3

Knauf, K., Memmert, D., & Brefeld, U. (2016). Spatio-temporal convolution kernels. *Machine learning*, *102*(2), 247–273. doi: 10.13140/RG.2.1.2572.9129

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, *28*(5), 1–26. Retrieved from `https://www.jstatsoft.org/v028/i05` doi: 10.18637/jss.v028.i05

Leser, R., Hoch, T., Tan, X., Moser, B., Kellermayr, G., & Baca, A. (2019). Finding efficient strategies in 3-versus-2 small-sided games of youth soccer players. *Kinesiology*, *51*(1.), 110–118. doi: 10.26582/k.51.1.7

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917. doi: 10.1287/isre.2013.0480

Link, D., & Hoernig, M. (2017). Individual ball possession in soccer. *PloS one*, *12*(7), e0179953. doi: 10.1371/journal.pone.0179953

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. doi: 10.1145/3236386.3241340

Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from opta sportsdata. *International Journal of Performance Analysis in Sport*, *13*(3), 803–821. doi: 10.1080/24748668.2013.11868690

Lopez, M. J., Matthews, G. J., & Baumer, B. S. (2018). How often does the best team win? a unified approach to understanding randomness in north american sport. *The Annals of Applied Statistics*, *12*(4), 2483–2516. doi: 10.48550/arXiv.1701.05976

MacLennan, T. (2005). Moneyball: The art of winning an unfair game. *Journal of Popular Culture*, *38*(4), 780.

Mehta, S., Phatak, A., Memmert, D., Kerruish, S., & Jamil, M. (2022). Seam or swing? identifying the most effective type of bowling variation for fast bowlers in men's international 50-over cricket. *Journal of Sports Sciences*, 1-5. doi: 10.1080/02640414.2022.2094140

Montoliu, R., Martín-Félez, R., Torres-Sospedra, J., & Martínez-Usó, A. (2015). Team activity recognition in association football using a bag-of-words-based method. *Human movement science*, *41*, 165–178. doi: 10.1016/j.humov.2015.03.007

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology*, *2*, 80–203.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: a review and a framework. *Measurement in Physical Education and Exercise Science*, *17*(3), 171–186. doi: 10.1080/1091367X.2013.805137

Oza, N. C., & Tumer, K. (2008). Classifier ensembles: Select real-world applications. *Information fusion*, *9*(1), 4–20. doi: 10.1016/j.inffus.2007.07.002

O'Donoghue, P. (2007). Reliability issues in performance analysis. *International Journal of Performance Analysis in Sport*, *7*(1), 35–48. doi: 10.1080/24748668.2007.11868386

Pappalardo, L., & Cintia, P. (2018). Quantifying the relation between performance and success in soccer. *Advances in Complex Systems*, *21*(03n04), 1750014. doi: 10.1142/S021952591750014X

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, *6*(1), 1–15. doi: 10.1038/s41597-019-0247-7

Paul, D. J., Bradley, P. S., & Nassis, G. P. (2015). Factors affecting match running performance of elite soccer players: shedding some light on the complexity. *International journal of sports physiology and performance*, *10*(4), 516–519. doi: 10.1123/IJSPP.2015-0029

Phatak, A., Mujumdar, U., Rein, R., Wunderlich, F., Garnica, M., & Memmert, D. (2020). Better with each throw—a study on calibration and warm-up decrement of real-time consecutive basketball free throws in elite nba athletes. *German Journal of Exercise and Sport Research*, *50*(2), 273–279. doi: 10.1007/s12662-020-00646-x

Phatak, A. A., Mehta, S., Wieland, F.-G., Jamil, M., Connor, M., Bassek, M., & Memmert, D. (2022). Context is key: normalization as a novel approach to sport specific preprocessing of kpi's for match analysis in soccer. *Scientific Reports*, *12*(1), 1–6. doi: 10.1038/s41598-022-05089-y

Phatak, A. A., Rein, R., & Memmert, D. (2021). The dirty league: English premier league provides higher incentives for fouling as compared to other european soccer leagues. *Journal of Human Kinetics*, *80*(1), 263-276. doi: 10.2478/hukin-2021-0095

Phatak, A. A., Wieland, F.-G., Vempala, K., Volkmar, F., & Memmert, D. (2021). Artificial intelligence based body sensor network framework—narrative review: Proposing an end-to-end framework using wearable sensors, real-time location systems and artificial intelligence/machine learning algorithms for data collection, data mining and knowledge discovery in sports and healthcare. *Sports medicine-open*, *7*(1), 1–15. doi: 10.1186/s40798-021-00372-0

Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (p. 1605–1613). New York, NY, USA: Association for Computing Machinery. Retrieved from `10.1145/3097983.3098051` doi: 10.1145/3097983.3098051

Raghupathi, W., et al. (2010). Data mining in health care. *Healthcare informatics: improving efficiency and productivity*, *211*, 223.

Raja, M. A. M., Manasa, V. V. L., Reddy, D. S. N., & Sundari, K. S. (2021). Applying data science for cricket predictions. *Annals of the Romanian Society for Cell Biology*, 1853–1863. Retrieved from `https://www.annalsofrscb.ro/index.php/journal/article/view/4713`

Rajšp, A., & Fister, I. (2020). A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences*, *10*(9), 3013. doi: 10.3390/app10093013

Razavi, A. R., Gill, H., Åhlfeldt, H., & Shahsavar, N. (2005). A data pre-processing method to increase efficiency and accuracy in data mining. In *Conference on artificial intelligence in medicine in europe* (pp. 434–443). doi: 10.1007/11527770_59

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532–538. doi: 10.1007/978-0-387-39940-9_565

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1–13. doi: 10.1186/s40064-016-3108-2

Rein, R., Raabe, D., & Memmert, D. (2017). "which pass is better?" novel approaches to assess passing effectiveness in elite soccer. *Human movement science*, *55*, 172–181. doi: 10.1016/j.humov.2017.07.010

Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F., & Baca, A. (2022). Machine learning application in soccer: A systematic review. *Biology of Sport*, *40*(1), 249–263. doi: 10.5114/biolsport.2023.112970

Robberechts, P., & Davis, J. (2020). How data availability affects the ability to learn good xg models. In *International workshop on machine learning and data mining for sports analytics* (pp. 17–27). doi: 10.1007/978-3-030-64912-8_2

Roy, R., Paul, A., Bhimjyani, P., Dey, N., Ganguly, D., Das, A. K., & Saha, S. (2020). A short review on applications of big data analytics. *Emerging Technology in Modelling and Graphics*, 265–278. doi: 10.1007/978-981-13-7403-6_25

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems (cts)* (p. 42-47). doi: 10.1109/CTS.2013.6567202

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, *37*(3), 297–336. doi: 10.1023/A:1007614523901

Severini, T. A. (2020). *Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports*. Chapman and Hall/CRC.

Sloan, M. (2017). *Sports analytics conference.* Retrieved from `https://www.sloansportsconference.com/` (Accessed = 2022-06-01)

Sun, X., Davis, J., Schulte, O., & Liu, G. (2020). Cracking the black box: Distilling deep sports analytics. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 3154–3162). doi: 10.1145/3394486.3403367

Swartz, T. B. (2020). Where should i publish my sports paper? *The American Statistician*, *74*(2), 103–108. doi: 10.1080/00031305.2018.1459842

Taborri, J., Keogh, J., Kos, A., Santuz, A., Umek, A., Urbanczyk, C., ... Rossi, S. (2020). Sport biomechanics applications using inertial, force, and emg sensors: A literature overview. *Applied bionics and biomechanics*, *2020*. doi: 10.1155/2020/2041549

Van Roy, M., Robberechts, P., Decroos, T., & Davis, J. (2020, Dec). Valuing on-the-ball actions in soccer: A critical comparison of xt and vaep. In *Proceedings of the AAAI-20 workshop on artifical intelligence in team sports.* AI in Team Sports Organising Committee. Retrieved from `https://lirias.kuleuven.be/2913207`

Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International journal of multimedia information retrieval*, *1*(3), 153–172. doi: 10.1007/s13735-012-0016-2

Wagenaar, M., Okafor, E., Frencken, W., & Wiering, M. A. (2017). Using deep convolutional neural networks to predict goal-scoring opportunities in soccer. In *International conference on pattern recognition applications and methods* (Vol. 2, pp. 448–455). doi: 10.5220/0006194804480455

Wasserstein, R. L., & Lazar, N. A. (2016). *The asa statement on p-values: context, process, and purpose* (Vol. 70) (No. 2). Taylor & Francis. doi: 10.1080/00031305.2016.1154108

Wunderlich, F., Seck, A., & Memmert, D. (2021). The influence of randomness on goals in football decreases over time. an empirical analysis of randomness involved in goal scoring in the english premier league. *Journal of Sports Sciences*, *39*(20), 2322–2337. doi: 10.1080/02640414.2021.1930685

Zago, M., Sforza, C., Dolci, C., Tarabini, M., & Galli, M. (2019). Use of machine learning and wearable sensors to predict energetics and kinematics of cutting maneuvers. *Sensors*, *19*(14), 3094. doi: 10.3390/s19143094

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists.* " O'Reilly Media, Inc.".